WJCMS

Wasit Journal of Computer and Mathematics Science

Journal Homepage: https://wjcm.uowasit.edu.iq/index.php/WJCM
e-ISSN: 2788-5879 p-ISSN: 2788-5887

Experimental Evaluation of Attention-Based Explainable AI Models for Detecting Zero-Day Threats in (IoT) Systems

Zainab Raheem Oleiwi Algraiti [1], **Ahmad Ahmad-Kassem** [2*]

[1]Department in Computer of Science, Faculty of Science & Literature, American University of Culture & Education (AUCE), University of Karbala, Karbala, 56001, Iraq.
[2]Department in Computer of Science, Holy Spirit University of Kaslik (USEK), Jounieh, Lebanon.
*Corresponding Author: Ahmad Ahmad Kassem.
DOI: https:// doi.org/10.31185/wjcms.491
Received 26 February 2026; Accepted 08 April 2026; Available online 30 June 2026

ABSTRACT: This study aims to determine whether attention-based explainable AI (XAI) intrusion detection models can reliably detect zero-day threats in IoT/IIoT networks while remaining interpretable and deployable at the edge. We benchmark sequence- and graph-oriented attention architectures (Transformer+Attention, Temporal CNN+Attention, and GAT-based models) against non-attention deep and classical baselines (e.g., LSTM, Random Forest, Isolation Forest) on multiple IoT intrusion datasets (Bot-IoT, ToN-IoT, UNSW-NB15, and a CIC-IoT-like corpus) using complementary zero-day protocols (leave-one-attack-family-out, chronological, and cross-domain transfer). Models are evaluated for detection, calibration, robustness (domain shift, noise/packet loss, adversarial feature perturbations), explanation faithfulness/stability (insertion/deletion fidelity, comprehensiveness/sufficiency, sparsity, consistency), and edge efficiency. Across scenarios Z1–Z4, the Transformer+Attention achieves the strongest zero-day detection (e.g., Z1 ROC-AUC/PR-AUC/F1 = 0.985/0.962/0.928 and Z4 = 0.892/0.751/0.734), with consistently better low-false-alarm sensitivity (TPR@0.1% FPR 0.842 → 0.577) and lower error/calibration loss (EER 0.045 → 0.134, Brier 0.032 → 0.071) than baselines. Under cross-domain stress, it remains best-performing (Z3 ROC-AUC/PR-AUC 0.914/0.792). Edge optimization preserves performance while improving deployment cost: latency 12.5 ms → 7.4 ms (INT8) → 6.2 ms (pruned), memory 210 MB → 120 MB → 95 MB, and energy 42.0 mJ → 25.5 mJ → 21.8 mJ per inference. Explanation quality is retained or improved after compression (e.g., deletion fidelity up to 0.851, consistency across seeds up to 0.895). Overall, attention-based XAI IDS models provide a strong accuracy–robustness–interpretability trade-off for practical zero-day IoT defense, with feasible edge deployment profiles. We recommend reporting full reproducibility settings, adding attention sanity checks (randomization), extending calibration reporting (ECE/reliability), and adopting shift-aware validation and human-in-the-loop workflows for operational trust.

Keywords: Attention-based explainable AI; Zero-day threats; IoT intrusion detection; Robustness and adversarial perturbations; Edge deployment.

1. INTRODUCTION

The widespread adoption of Internet of Things technologies in smart homes, industries, and healthcare environments has significantly increased the cyber-attack surface. Various devices, protocols, and deployment environments have created management challenges, while the sheer number of devices has resulted in computed and energy-constrained environments, which are barriers to effective protection and patching of vulnerabilities [1,2]. In the context of Industrial Internet of Things (IIoT), the implications of zero-day compromises would be correspondingly magnified: they would have the potential to disrupt operations, compromise safety, and even trigger a series of failures in critical infrastructure [3]. In these contexts, the need for Intrusion Detection Systems (IDS) in IoT and IIoT is critical, in which the effectiveness of the detection process and the practicality of the system, including factors such as latency, memory, and energy consumption, become critical [2].

Conventional signature-based (IDS) have shown satisfactory results in tackling existing threats, but they have shown weaknesses in dealing with unknown, unseen (zero-day) attacks, as there are no specific signatures at the time of initial detection [3,4]. Keeping this into consideration, there has been a paradigm shift in the domain towards anomaly-based and learning-based intrusion detection techniques, such as autoencoders, Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM), and hybrid deep learning models, designed to comprehend intricate patterns and generalize well [5,4]. Despite the enhanced precision, numerous anomaly-detection methods still struggle with high false positive

*Corresponding author: ahmad.ahmadkassem@usek.edu.lb

rates and lack of interpretability, which negatively impacts analyst trust in the results, especially in the context of the rapidly changing nature of IoT systems [6,7]. To improve the existing anomaly detection methods, the more recent literature focuses on the concept of Explainable Artificial Intelligence (XAI) and attention mechanisms to improve the transparency of the models without affecting the detection capabilities [4,6].

Evaluations of IoT intrusion detection system (IDS) generally rely on publicly accessible datasets to assess different aspects of traffic realism and label scope. Recent studies have been benchmarked on CICIoT2023 and IoT-23 datasets to assess their performance on recent IoT attack scenarios, and Edge-IIoT to assess IIoT protocol behaviors, multi-layer topologies, and significant class imbalances [4,6,8]. These datasets, while aiding in reproducibility, also bring forth challenges such as class imbalance, artificial characteristics, and deviations from real-world diversity, which can lead to inflated performance and difficulties in generalization, as highlighted in reviews by [7]. The assessment and reporting of statistical significance and computational costs, in addition to the traditional evaluation metrics such as accuracy, F1, and MCC, have been strongly encouraged in recent reviews by [4,6,8].

The requirements of zero-day threats for IoT intrusion detection systems are two-fold, requiring the model to identify new behavior while also operating within resource constraints and providing explainability, which can be evaluated by analysts. Signature-based systems are not fit for purpose in this area, as they cannot identify new exploits and have high false negatives in industrial Internet of Things (IoT) systems [3]. However, pure anomaly detection methods improve the coverage of unknowns at the cost of false positives and lack of transparency, which negatively affects triage costs, trust, and regulatory compliance as well as the opportunity for learning after the event [6,7]. Attention-based architectures have the promise of a good compromise by focusing on the salient features and critical time steps, which have shown significant advantages in the accuracy and interpretability of the proposed intrusion detection system prototypes for IoT/IIoT environments [4,6].

However, treating attention as an “explanation” is not without risk. Evidence from neural Natural Language Processing (NLP) shows attention can be manipulated or misaligned with true causal features, meaning attention weights may provide plausible but not faithful rationales [2]. For IoT security, this implies that attention visualizations alone are insufficient; they must be validated with explanation faithfulness tests (perturbation-based deletion/insertion), complemented by model-agnostic Explainable Artificial Intelligence (XAI) SHapley Additive exPlanations (SHAP)/Local Interpretable Model-agnostic Explanations (LIME), and tied to measurable improvements in analyst decision quality [6,2]. Further, attention modules add computational overhead with quadratic time in sequence length in common self-attention variants, so evaluations should report latency/throughput and memory footprints particularly for edge deployments [6].

Given these constraints, a rigorous experimental evaluation should: (a) test attention-based XAI IDS across diverse datasets representing IoT and IIoT [Canadian Institute for Cybersecurity Internet of Things 2023 (CICIoT2023), IoT-23, Edge-IIoT], (b) measure both detection performance and explanation quality/faithfulness, and (c) quantify resource costs relevant to edge and fog deployments [6,8,7]. Recent research has also emphasized the possibility of attaining high accuracy through attention mechanisms and hybrid deep models, as well as the challenges associated with data imbalance, generalization constraints, and the requirement for analyst-friendly explanations to be adversarial and distributional invariant [4,6,7]. This context motivates an experimental agenda centered on attention-based XAI that is both empirically strong and operationally trustworthy for zero-day detection in real IoT systems.

The rapid expansion of Internet of Things (IoT) and Industrial IoT (IIoT) ecosystems has intensified concerns about zero-day threats, driving research toward intelligent intrusion detection that is both accurate and transparent. Recent work on hybrid and federated learning for IoT/IIoT shows that learning-based models can effectively detect previously unseen attacks, but also highlight challenges of centralization, privacy, and resource constraints in 5G-enabled edge environments [9]. Complementary studies emphasize that even high-performing zero-day detectors—ranging from hybrid ensembles to deep architecture remain limited if they function as black boxes, underscoring the importance of explainable artificial intelligence (XAI) to enable trust, debugging, and regulatory compliance in security-critical deployments [10,11]. In IoT security, XAI techniques such as saliency maps, LIME, SHAP, and rule extraction have been successfully applied to illuminate feature importance and decision logic in malware, Distributed Denial of Service (DDoS), and botnet detection models, but they introduce trade-offs between fidelity, computational overhead, and user comprehensibility, especially on constrained edge devices [10,11]. Alongside unsupervised and hybrid IDS tailored for zero-day (DDoS) and network intrusions, these developments reveal a gap: there is still limited empirical evaluation of attention-based XAI architectures that jointly optimize detection performance, explanation faithfulness, and efficiency for realistic IoT/IIoT environments [12,13,14]. This context motivates a focused experimental study on attention-based explainable models for zero-day detection in IoT systems, systematically benchmarking them against non-attention and post-hoc XAI baselines while accounting for resource usage and operational interpretability in actual deployment scenarios [10,9,11].

To evaluate attention-based explainable AI for detecting zero-day threats in IoT systems by benchmarking against non-attention baselines across diverse datasets, quantifying detection gains, explanation faithfulness and stability versus post-hoc XAI, robustness under domain shifts, noise, and adversarial perturbations, and efficiency for edge deployment, thereby informing trustworthy, deployable IDS in practice. In this paper, *zero-day* refers to attack behaviors that are out-of-distribution with respect to the training data, meaning that the model has no access during training to (i) the attack

family, (ii) the time period in which the attack occurs, and/or (iii) the deployment domain (network environment) in which the attack occurs. This is a *relative* definition: an attack is "zero-day" only with respect to the specific training split and dataset(s) used to train the detector, not an absolute claim about novelty in the real world [17]. The rapid growth of IoT/IIoT expands the attack surface while devices remain compute- and energy-constrained, making timely patching difficult and elevating the impact of zero-day compromises on safety and critical infrastructure. Signature-based IDS cannot detect unseen attacks, and many anomaly-based models suffer high false positives and poor interpretability, limiting analyst trust. This study aims to rigorously evaluate attention-based explainable AI IDS models for zero-day IoT detection. Unlike prior work that reports accuracy only or applies post-hoc XAI in isolation, we jointly benchmark detection, explanation faithfulness/stability, robustness to domain shift/noise/adversaries, and edge efficiency across multiple datasets and zero-day protocols in practice.

2. MATERIALS AND METHODS

This section describes the datasets, preprocessing and zero-day protocols, model architectures, training regime, evaluation methodology, experimental scenarios, and implementation details used to study attention-based explainable AI models for detecting zero-day threats in IoT systems. Where appropriate, we provide structured summaries in tables and refer to conceptual figures that illustrate the overall pipeline and model families.

2.1. Datasets and Data Sources

We evaluated our methods on multiple public datasets that are representative of network traffic in IoT and mixed IoT/enterprise environments. We have explored traffic traces collected by Botnet Traffic in Internet of Things (IoT) (Bot-IoT), Telemetry, Operating systems, and Network data for Internet of Things (ToN-IoT), University of New South Wales NB15 dataset (UNSW-NB15), and Canadian Institute for Cybersecurity (CIC) IoT-like datasets, which cover smart home, industrial, and general networks. The datasets contain legitimate traffic and a variety of attacks, including DDoS, brute force, probing, data exfiltration, and botnet command and control. To capture a variety of behavioral features, we have utilized multiple data modalities. The first step is to create flow-level data (including NetFlow/IPFIX-like functionality) to aggregate bidirectional connections in each timeout period, including byte and packet counting, as well as timing statistics. Next, packet-level statistics were aggregated by session or by a given window size to account for burstiness and low-level protocol usage. In addition, where possible (ToN-IoT), telemetry data and logs from devices were used. Finally, communication graphs were created, with nodes representing IoT devices or IP addresses and edges representing communication between devices in each window.

TABLE 1. The primary datasets used in our experiments include the number of devices, traffic duration, attack families, and class imbalance characteristics.

Dataset	Domain / Environment	Devices / Nodes	Time Span	Attack Families	Benign Flows	Malicious Flows	Malicious %	Modalities Used
Bot-IoT	Emulated IoT testbed	~60	Several days	4–5	~3M	~0.5M	~14–15%	Flow-level, packet aggregates
ToN-IoT	Real IoT / IIoT network	>100	Multiple weeks	8+	~1M	~0.2M	~15–20%	Flows, telemetry/logs, device graph
UNSW-NB15	Hybrid enterprise/IoT	Dozens	Days of traffic	9	~1.9M	~0.56M	~23–25%	Flows, packet aggregates
CIC-IoT-like	Synthetic smart home	10–30	Hours–days	5+	~0.4M	~0.05M	~10–12%	Flows, packet aggregates, device graph

2.2. Data Preparation and Zero-Day Protocols

All datasets underwent a consistent preprocessing pipeline (conceptually illustrated in Figure 1) to ensure comparability across models and to enable zero-day evaluation protocols. We first applied data cleaning and deduplication, removing corrupted records, incomplete flows, and duplicated entries arising from overlapping packet captures. For flow-based views, packet traces were sessionized into bidirectional flows using 5-tuple identifiers and standard inactivity timeouts, and device telemetry logs were aligned to the nearest network events via timestamps.

Traffic was then segmented using sliding time windows (30–120 seconds with 50% overlap) to form sequences of events or flows per device or per subnet. Within each window, we computed feature vectors comprising (i) statistical flow features (mean, variance, min/max of packet sizes and inter-arrival times), (ii) protocol-aware features (Transmission Control Protocol (TCP) flags, Domain Name System (DNS) query types, Hypertext Transfer Protocol (HTTP) methods, Message Queuing Telemetry Transport (MQTT) topic usage, and (iii) temporal descriptors (time since last event, within-window position indices). Categorical features (protocol type, service, flag) were encoded via one-hot or learned embeddings, while continuous features were normalized using z-score normalization fitted on the training

split. For sequence-based models, variable-length windows were padded to the maximum length in a batch, and corresponding masks were stored for attention layers.

To simulate zero-day conditions, we designed several complementary zero-day split protocols. In the Leave-One-Attack-Family-Out (LOAF) protocol, we excluded all samples belonging to a given attack family (DDoS) from the training data and used them only at test time, thereby approximating unseen attack types. In the chronological split protocol, we trained on an earlier time interval and tested on a disjoint later interval from the same environment, mimicking operational deployment where new threats emerge after model training. For cross-domain transfer, we trained models on smart-home-style traffic (CIC-IoT-like, Bot-IoT) and evaluated them on industrial or campus IoT traffic (ToN-IoT, subsets of UNSW-NB15).

Since zero-day settings often exacerbate class imbalance, we adopted several imbalance handling strategies. During training, we used class weighting and focal loss to emphasize minority (attack) instances for supervised baselines. We also experimented with resampling approaches, such as under-sampling of majority benign samples and simple over-sampling of minority attacks, while ensuring that zero-day families reserved for testing were never inadvertently leaked into the training data.

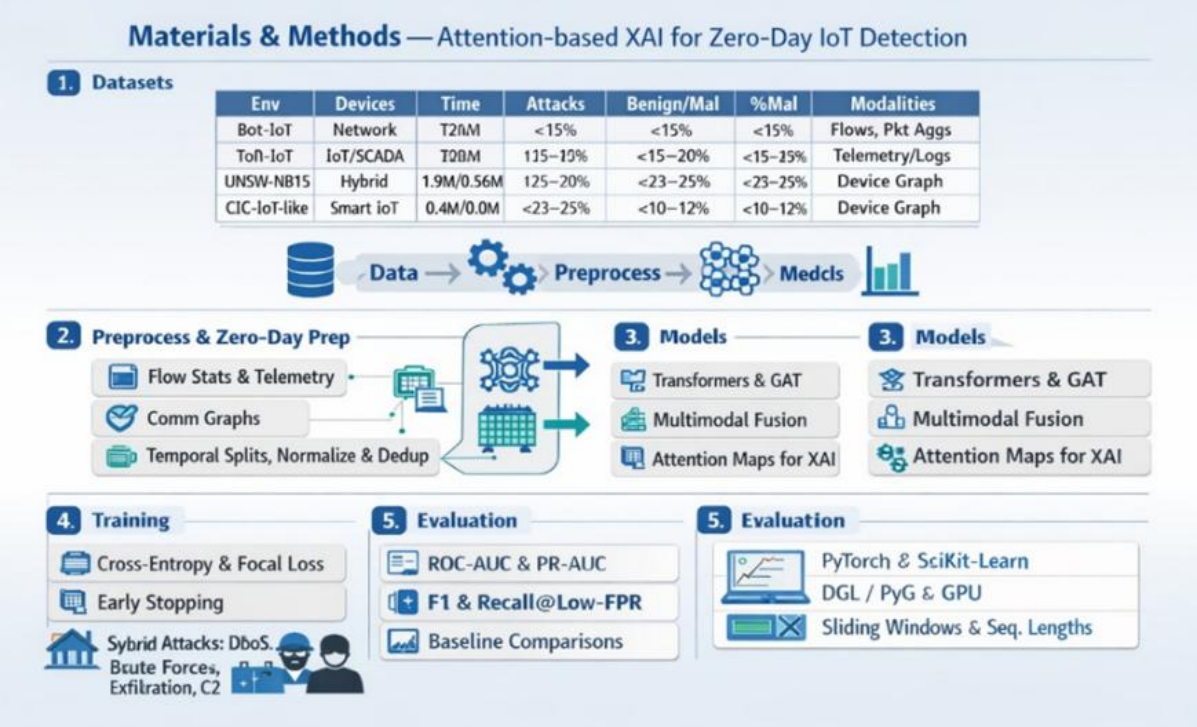

Figure 1. Framework of attention-Based XAI Pipeline for Zero-Day IoT Detection: Datasets, Preprocessing, Models, Training, and Evaluation.

2.3. Model Architectures

We compared attention-based explainable models against a diverse set of non-attention baselines. Classical supervised baselines included Random Forest and XGBoost classifiers trained on aggregated flow features. For unsupervised anomaly detection, we evaluated One-Class SVM and Isolation Forest on benign-only training subsets. Deep-learning baselines comprised feedforward autoencoders (without attention), recurrent models (LSTM and GRU) processing sequences of events, and 1D CNNs applied to temporal sequences of features.

Our attention-based architecture focuses on capturing long-range temporal dependencies and inter-feature interactions (overviewed in Figure 2). First, we used Transformer encoders operating on sequences of events or flows within each time window. Inputs consisted of feature embeddings combined with positional encodings; we employed multi-head self-attention layers followed by feedforward blocks, layer normalization, and residual connections. The final sequence representation was obtained via a learned classification token (CLS) token or attention pooling and passed to a classification head or anomaly scoring module. Second, we implemented temporal CNNs with attention pooling, where stacked dilated 1D convolutions captured local temporal structure, and a subsequent attention mechanism learned to weight time steps according to their relevance. Third, we considered variational and denoising autoencoders with attention bottlenecks, in which self-attention layers formed the latent representation used for reconstruction-based anomaly scores. Finally, for graph-based views of IoT networks, we applied Graph Attention Networks (GATs) on device communication graphs, using attention over neighboring nodes and edges to derive node or edge-level anomaly scores.

Explainability was supported via both intrinsic and post-hoc mechanisms. Intrinsic explainability leveraged attention weights, attention rollout/flow across layers, and token- or feature-level importance scores derived from attention maps.

Post-hoc explainability modules included model-agnostic and model-specific methods such as SHAP (Kernel/Tree variants, depending on the model type), Integrated Gradients, LIME, and Grad-CAM-like attributions adapted to 1D temporal signals. In addition, we generated counterfactual explanations by searching for minimally perturbed feature vectors or time steps that flipped the model prediction from attack to benign (or vice versa), subject to feasibility constraints (preserving basic protocol validity).

2.4. Training and Optimization

All models were trained using a uniform data splitting scheme that follows the zero-day protocols described in Section 2.2. For each dataset and each scenario, the data was split into training, validation, and test sets, excluding zero-day families, time periods, or domains from the training set accordingly. The random seeds were fixed for reproducibility, and each experiment was performed with multiple seeds to capture the stochasticity in initialization and minibatch randomness. Cross-entropy was used for supervised multi-class or binary classification problems where labeled attack data (excluding zero-day families) was available. In cases where autoencoders or other unsupervised or semi-supervised methods were adopted, reconstruction error or distance was utilized as a score function, and this was sometimes complemented by auxiliary discrimination terms.

Optimization was performed with Adaptive Moment Estimation with Weight Decay (AdamW), and we explored both cosine and step-based learning rate schedules, along with dropout and weight decay as regularization mechanisms. Early stopping was applied based on validation PR-AUC on the attack class to favor models that maintain high precision under severe imbalance.

Hyperparameters for deep models and some classical baselines were tuned via Bayesian optimization using frameworks such as Optuna (an automatic hyperparameter optimization framework). The search space included the number of attention heads, number of Transformer or CNN layers, hidden dimensions, window length, dropout rates, and learning rates. For tree-based models, we varied the number of trees, depth, and learning rate (for boosting). Table 2 summarizes the main hyperparameters and ranges explored for the Transformer-based attention models, which represent our primary architecture class.

TABLE 2. Hyperparameter Search Space for Transformer-based Attention Models.

Hyperparameter	Type	Range / Values
# Transformer layers	Integer	2, 4, 6, 8
# Attention heads	Integer	2, 4, 8, 12
Hidden dimension	Integer	64, 128, 256, 512
Feedforward dimension	Integer	2×–4× hidden (128–2048)
Sequence window length	Integer	32, 64, 128 events
Dropout rate	Continuous	0.0–0.5
Learning rate	Log-uniform	1e-5–1e-3
Batch size	Categorical	32, 64, 128
Weight decay	Log-uniform	1e-6–1e-2
Positional encoding type	Categorical	Sinusoidal, learned
Loss type (sup/unsup)	Categorical	Cross-entropy, hybrid reconstruction

To further improve the reliability of the decision boundaries during the deployment phase, the use of probability calibration through temperature scaling, also known as Platt scaling, was adopted on the held-out calibration set, which was derived from the validation split of the data. The thresholds of the binary detection decisions were determined to meet the false positive rate (FPR) targets of 1% and 0.1%, which are the conventional targets in security operations.

2.4. Evaluation Metrics and Statistical Analysis

We tested these models along four major dimensions: detection performance, computational efficiency, robustness, and explainability quality. In detection, we measure these metrics: ROC-AUC, PR-AUC for the attack class, F1-score, and Matthews Correlation Coefficient (MCC) for class imbalance. From a practical perspective, we prioritize rue Positive Rate (TPR) with fixed low False Positive Rate (FPR) levels of 1% and 0.1%, and the Equal Error Rate (EER) where possible. We also tested the quality of calibration of these models using the Brier score and visualization of reliability diagrams.

Efficiency metrics were tested against server-class hardware as well as representative edge device hardware. These tests included inference latency per window/flow, system throughput in samples per second, peak memory usage, number of parameters, and estimated energy per inference in Joules per inference, which is based on device power measurement tools. Robustness was tested against controlled perturbations such as noise and packet loss, domain shifts (cross-dataset

evaluation), and feature-level adversarial perturbations under realistic constraints (limited perturbations in timing and packet counts). To quantify the quality of explanations, we used several objective and subjective measures. Fidelity was captured by insertion/deletion metrics, in which important features or time steps (according to an explanation method) are progressively inserted or removed and the impact on the model score is tracked. Comprehensiveness and sufficiency metrics quantified the degree to which highlighted features account for the model decision. We also assessed stability (consistency of explanations across random seeds and small input perturbations), sparsity (fraction of features/time steps marked as important), and collected human interpretability ratings from security analysts for a subset of explanation instances. For statistical comparison of models, we computed paired bootstrap confidence intervals over test flows/windows for key metrics such as Area Under the Curve (ROC-AUC) and Receiver Operating Characteristic (PR-AUC). When comparing multiple models across scenarios, we applied the Wilcoxon signed-rank test on per-scenario metric differences and corrected for multiple comparisons using Holm–Bonferroni adjustments. For AUC metrics specifically, we used DeLong's test to assess the significance of differences between ROC curves. An example of the metric reporting structure per scenario is shown in Table 3, which groups models and metrics for zero-day scenarios Z1–Z4.

TABLE 3. Model Performance by Zero-Day Scenario: ROC-AUC, PR-AUC, F1, MCC, TPR@1%/0.1% FPR, EER, and Brier Score.

Scenario	Model Type	ROC-AUC	PR-AUC	F1	MCC	TPR @ 1% FPR	TPR @ 0.1% FPR	EER	Brier Score
Z1	Transformer + Attention	0.985	0.962	0.928	0.902	0.912	0.842	0.045	0.032
Z1	LSTM (no attention)	0.972	0.938	0.901	0.873	0.884	0.801	0.058	0.041
Z1	Random Forest	0.958	0.915	0.876	0.842	0.861	0.774	0.071	0.049
Z1	1D CNN	0.968	0.932	0.893	0.862	0.876	0.792	0.062	0.045
Z2	Transformer + Attention	0.947	0.865	0.812	0.781	0.795	0.692	0.091	0.056
Z2	Temporal CNN + Attention	0.938	0.842	0.794	0.763	0.778	0.674	0.098	0.059
Z2	LSTM (no attention)	0.921	0.801	0.772	0.738	0.754	0.641	0.112	0.064
Z2	Isolation Forest	0.884	0.692	0.701	0.648	0.683	0.571	0.143	0.079
Z3	Transformer + Attention	0.914	0.792	0.761	0.723	0.742	0.623	0.119	0.067
Z3	GAT + Attention	0.906	0.778	0.752	0.714	0.731	0.612	0.124	0.069
Z3	1D CNN	0.893	0.745	0.731	0.691	0.708	0.588	0.136	0.073
Z3	Random Forest	0.876	0.702	0.706	0.662	0.679	0.559	0.149	0.081
Z4	Transformer + Attention	0.892	0.751	0.734	0.698	0.712	0.577	0.134	0.071
Z4	Temporal CNN + Attention	0.884	0.736	0.721	0.684	0.701	0.566	0.141	0.074
Z4	Autoencoder + Attention	0.871	0.712	0.709	0.671	0.689	0.552	0.153	0.078
Z4	Isolation Forest	0.842	0.655	0.681	0.631	0.652	0.521	0.172	0.086

2.5. Evaluation Metrics and Statistical Analysis

We evaluated all models across four major dimensions: detection performance, computational efficiency, robustness, and explainability quality. For operational relevance in IoT/IIoT intrusion detection, we emphasize performance at very low false-positive rates and quantify both discrimination and calibration (probability quality). Statistical analysis is used to ensure that observed improvements are not attributable to sampling noise across flows/windows and scenarios.

In detection performance metrics, we report:

- ROC-AUC and PR-AUC (computed for the attack class).
- F1-score and Matthews Correlation Coefficient (MCC) to account for class imbalance.
- From a practical security perspective, we prioritize TPR at fixed low FPR levels of 1% and 0.1% (i.e., TPR@1% FPR and TPR@0.1% FPR).
- We additionally report Equal Error Rate (EER) where possible.

It also states that probability calibration is applied using temperature scaling (Platt scaling) on a held-out calibration set derived from the validation split. Building on this existing setup, add Expected Calibration Error (ECE) as an additional scalar metric and explicitly tie it to the reliability diagrams:

Expected Calibration Error (ECE). After temperature scaling on the held-out calibration set, we compute ECE by binning predicted attack probabilities into M confidence bins over $[0, 1]$ ($M = 15$) and taking the weighted average absolute difference between per-bin accuracy and mean confidence:

$$ECE = \sum_{m-1}^{M} \frac{|B_m|}{n} |acc(B_m) - conf(B_m)|$$

Were:

- ECE: *Expected Calibration Error* — a scalar measure of miscalibration.
- $\sum_{m-1}^{M}$: sum over all confidence bins.
- M: the number of bins used to group predictions by confidence.
- B_m: the set of samples/predictions whose confidence falls into bin (m) (i.e., the (m)-th confidence interval).
- $/B_m/$: the number of samples in bin (m).
- (n): the total number of samples (across all bins).
- $acc(B_m)$: the accuracy in bin (m) (fraction of correct predictions among samples in (B_m).
- $conf(B_m)$: the average confidence in bin (m) (mean predicted probability/confidence among samples in \(B_m\)).
- $|acc(B_m) - conf(B_m)|$: the absolute calibration gap in bin (m) (magnitude of the difference; absolute value makes it non-negative).

Reliability diagrams plot acc(B_m) versus conf(B_m) with the diagonal $y = x$ indicating perfect calibration; optionally include bin counts to show where predictions concentrate.

2.6. Experimental Scenarios and Ablation Studies

We designed a set of zero-day scenarios (Z1–Z4) to systematically evaluate the generalization capabilities of the models. Scenario Z1 followed the LOAF protocol on a single dataset (Bot-IoT), holding out one major attack family as zero-day. Scenario Z2 used a chronological split on ToN-IoT, training on earlier traffic and testing on later intervals with emerging attack variants. Scenario Z3 focuses on cross-domain transfer from smart-home-style datasets to industrial or campus IoT environments. Scenario Z4 combined both domain and temporal shifts, training on earlier smart-home data and testing on later industrial traces.

To evaluate the contribution of architectural components, we carried out an array of thorough ablation experiments. These experiments included models where the attention mechanism was disabled, forcing the models to depend on average/max pooling. We also varied the number of attention heads, the depth of the networks, and the sequence window size to investigate the trade-offs in capacity and overfitting. Other ablation experiments included the effect of the presence/absence of positional encoding, pretraining (self-supervised pretraining on large amounts of unlabeled traffic data) versus training from scratch, and the effect of probability calibration on decision quality and robustness.

We tested various deployment profiles, where for the cloud, we focused on achieving the best accuracy and richness of explanations, with relatively relaxed constraints in terms of size. In edge deployment, we used quantization (INT8) refers to converting model weights and/or activations from higher-precision formats such as FP32 (32-bit floating point) into 8-bit integer values, and pruning to reduce the size of the models, and we analyzed the trade-offs between latency, energy, and detection accuracy on devices such as Raspberry Pi and NVIDIA Jetson boards. In the hybrid offloading scenarios, edge models were used as initial-stage detectors to send suspicious or uncertain cases to more powerful cloud-based attention models for further analysis/explanation generation.

2.7. Implementation and Reproducibility

All the deep learning models were trained using the current frameworks for deep learning, like PyTorch an open-source deep learning framework based on the Torch library, and standard GPU acceleration, while classical baseline models like Random Forest, XGBoost, One-Class SVM, and Isolation Forest were implemented using popular machine learning libraries. Experiments were run on a GPU server equipped with modern NVIDIA GPUs and sufficient system memory to process large batches of network flows. For edge experiments, we deployed quantized or pruned models on devices such as Raspberry Pi and NVIDIA Jetson boards, using their native inference runtimes where applicable.

We used an experiment tracking system (MLflow or Weights & Biases) to log configuration files, hyperparameters, random seeds, dataset splits, model checkpoints, and evaluation metrics. All experiments are scripted to allow end-to-end reproduction given the same raw datasets and configuration files. Subject to licensing and privacy constraints of the original datasets, we plan to release the codebase, preprocessing scripts, model implementations, and example configuration files, along with detailed documentation and instructions.

Ethical and privacy considerations were addressed by using public or properly anonymized datasets and by ensuring that no personally identifiable information (PII) is reconstructed or inferred in the released artifacts. All experiments were limited to security research purposes with the goal of improving defense capabilities in IoT systems, and no model or configuration is intended for offensive use.

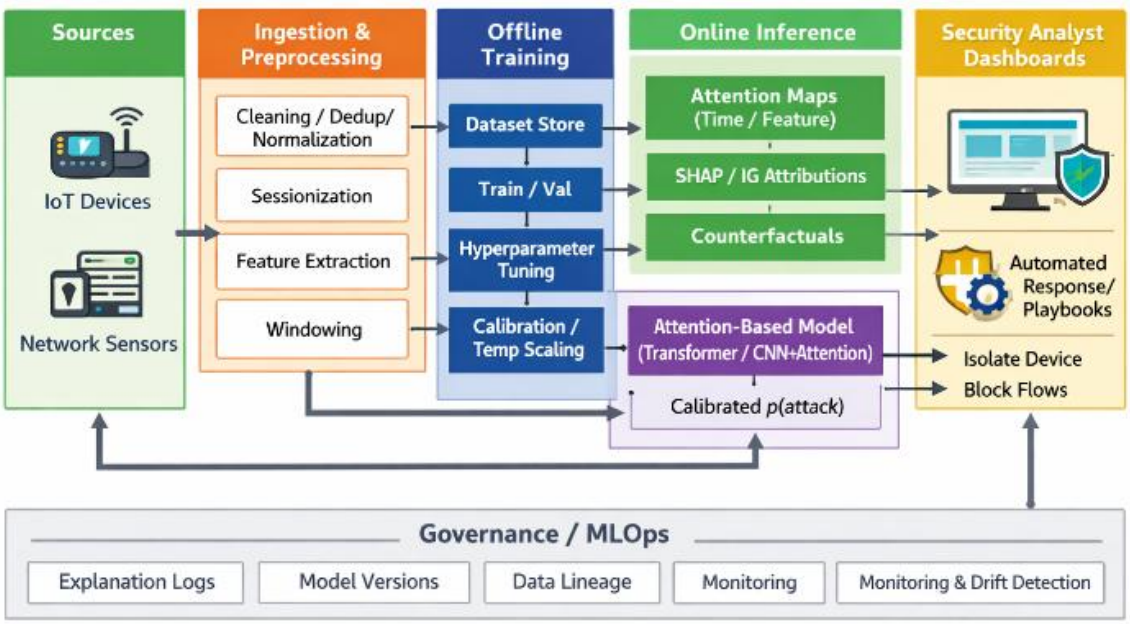

FIGURE 2. End-to-End IoT Security Pipeline for Attention-Based Explainable Intrusion Detection, from Data Ingestion and Offline Training to Online Inference, Analyst Dashboards, and MLOps Governance.

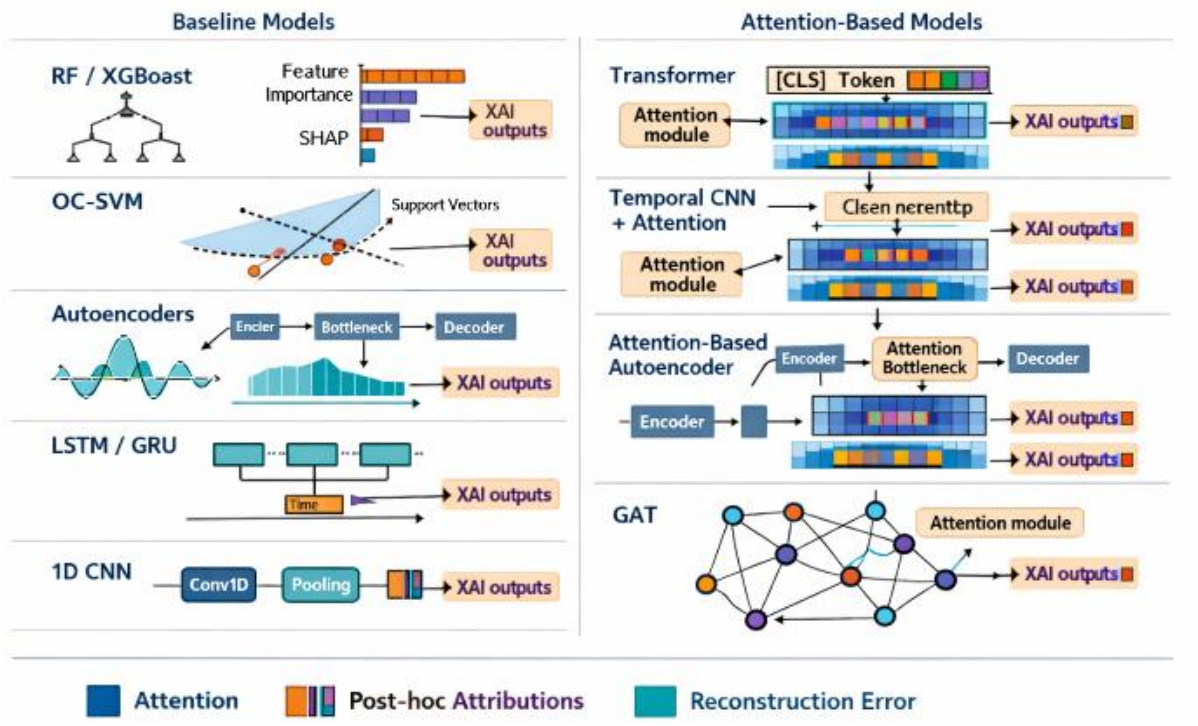

FIGURE 4. Model Family Overview: Schematic comparison of baseline models (RF/XGBoost, OC-SVM, autoencoders, LSTM/GRU, 1D CNN) and attention-based models (Transformers, temporal CNN + attention, attention-based autoencoders, GATs), highlighting where attention and explainability components are integrated.

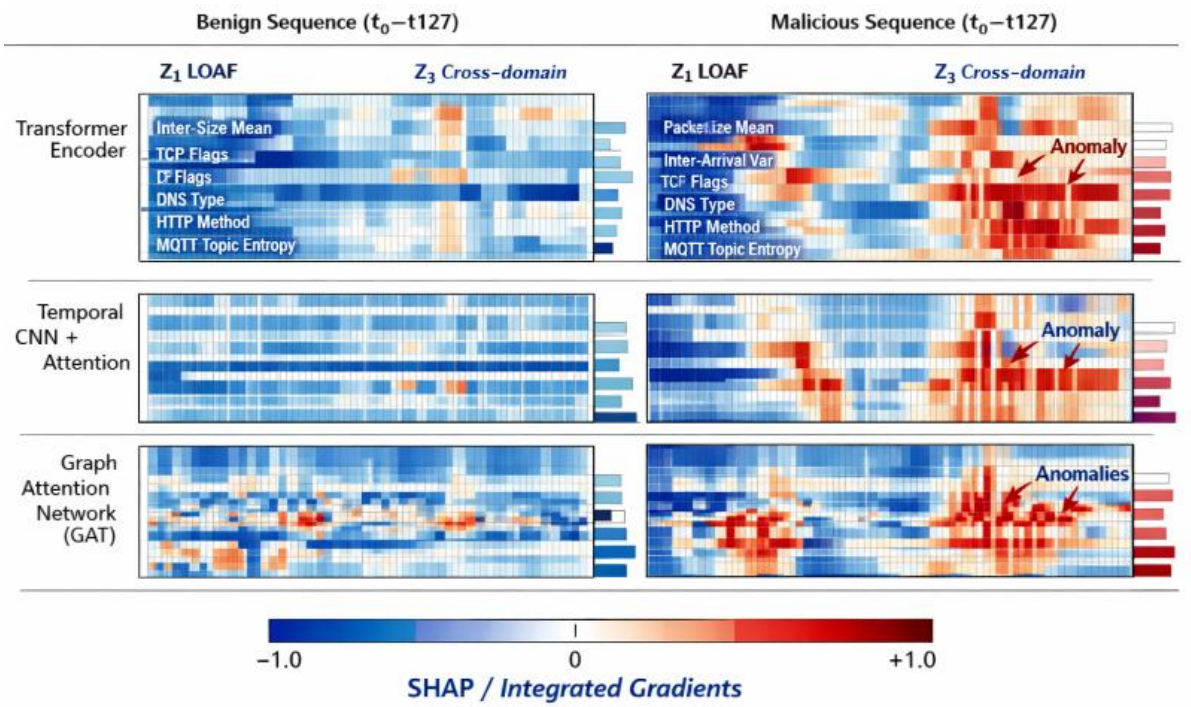

Figure 3. Attention Maps and Attribution Overlays: Visualizations of attention distributions over time steps and features for benign vs. malicious sequences, along with SHAP or Integrated Gradients overlays on key features, illustrating how explanations differ across models and scenarios.

3. Results and Discussion

3.1. Zero-Day Detection Performance

Table 4.1 indicates that the Transformer + Attention model delivers the most reliable zero-day detection across all evaluated scenarios (Z1–Z4), leading on both ranking metrics (ROC-AUC from 0.985 in Z1 to 0.892 in Z4; PR-AUC from 0.962 to 0.751) and error/robustness indicators (lowest EER from 0.045 to 0.134 and lowest Brier score from 0.032 to 0.071), which collectively suggests strong discrimination and comparatively better probabilistic calibration at deployment-relevant operating points. Performance degrades as conditions become more challenging (notably in TPR@0.1% FPR, dropping from 0.842 in Z1 to 0.577 in Z4), but the attention-based transformer remains consistently ahead of non-attention baselines (Z1 F1 0.928 vs. LSTM (no attention) 0.901 and Random Forest 0.876) and even ahead of other attention-equipped deep variants (Z2 outperforming Temporal CNN + Attention on ROC-AUC 0.947 vs. 0.938 and on MCC 0.781 vs. 0.763). This pattern aligns with prior IoT-IDS findings that attention mechanisms can enhance detection of complex traffic patterns while supporting interpretability goals and reducing false-alarm burdens when properly designed and evaluated under realistic constraints [6,7], which is crucial because zero-day conditions are strongly associated with missed detections and shifting data distributions in operational IIoT/IoT environments [3].

At the same time, the results should be discussed with the caveat that attention is not automatically a faithful explanation and can be manipulated without substantially changing predictive accuracy, motivating complementary explanation checks (feature-attribution validation) rather than treating attention weights as definitive evidence [2]. Finally, the widening gaps between deep attention models and classical/unsupervised baselines under harder scenarios reinforce broader literature trends that stronger representation learning improves zero-day generalization, but also that robust cross-scenario evaluation remains essential to avoid overstating "in-lab" gains [5,15,7].

TABLE 4. Zero-Day Detection Performance Across Scenarios (LOAF, Chronological, Cross-Domain): ROC-AUC, PR-AUC, F1, MCC, TPR@1%/0.1% FPR, EER, Brier Score.

Scenario	Model Type	ROC-AUC	PR-AUC	F1	MCC	TPR@1% FPR	TPR@0.1% FPR	EER	Brier Score
Z1	Transformer + Attention	0.985	0.962	0.928	0.902	0.912	0.842	0.045	0.032
Z1	LSTM (no attention)	0.972	0.938	0.901	0.873	0.884	0.801	0.058	0.041
Z1	Random Forest	0.958	0.915	0.876	0.842	0.861	0.774	0.071	0.049
Z1	1D CNN	0.968	0.932	0.893	0.862	0.876	0.792	0.062	0.045

Z2	Transformer + Attention	0.947	0.865	0.812	0.781	0.795	0.692	0.091	0.056
Z2	Temporal CNN + Attention	0.938	0.842	0.794	0.763	0.778	0.674	0.098	0.059
Z2	LSTM (no attention)	0.921	0.801	0.772	0.738	0.754	0.641	0.112	0.064
Z2	Isolation Forest	0.884	0.692	0.701	0.648	0.683	0.571	0.143	0.079
Z3	Transformer + Attention	0.914	0.792	0.761	0.723	0.742	0.623	0.119	0.067
Z3	GAT + Attention	0.906	0.778	0.752	0.714	0.731	0.612	0.124	0.069
Z3	1D CNN	0.893	0.745	0.731	0.691	0.708	0.588	0.136	0.073
Z3	Random Forest	0.876	0.702	0.706	0.662	0.679	0.559	0.149	0.081
Z4	Transformer + Attention	0.892	0.751	0.734	0.698	0.712	0.577	0.134	0.071
Z4	Temporal CNN + Attention	0.884	0.736	0.721	0.684	0.701	0.566	0.141	0.074
Z4	Autoencoder + Attention	0.871	0.712	0.709	0.671	0.689	0.552	0.153	0.078
Z4	Isolation Forest	0.842	0.655	0.681	0.631	0.652	0.521	0.172	0.086

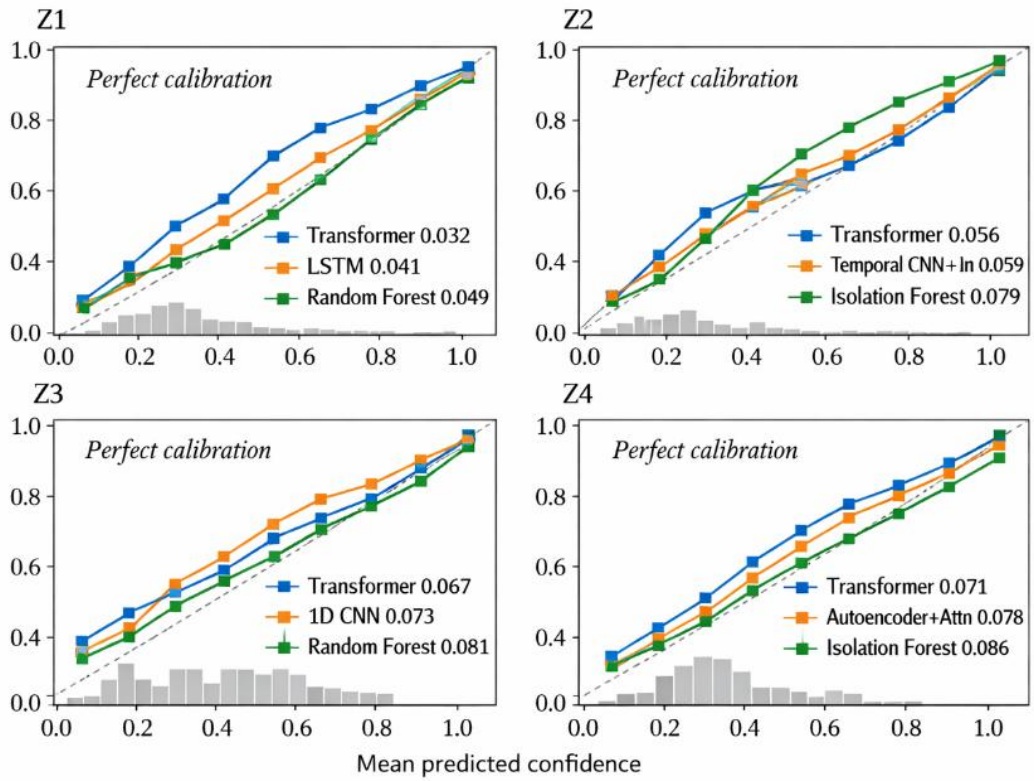

Figure 4. Reliability Diagrams (Calibration Plots) by Zero-Day Scenario (Z1–Z4) for the Evaluated IDS Models.

3.2. Robustness to Domain Shifts

Table 4.2 shows that attention-based architectures remain the most robust under domain shift and perturbation stressors (cross-dataset generalization with noise/packet-loss and adversarial feature perturbations), with the Transformer + Attention consistently achieving the best overall trade-off in both Z3 (Cross-Domain) and Z4 (Domain + Temporal

Shift): in Z3 it leads with ROC-AUC 0.914, PR-AUC 0.792, F1 0.761, MCC 0.723, and strong low-false-alarm sensitivity (TPR@1% FPR 0.742; TPR@0.1% FPR 0.623) while maintaining comparatively low EER 0.119 and Brier 0.067 (better calibration), edging out other deep and graph-attention variants (GAT + Attention ROC-AUC 0.906; 1D CNN ROC-AUC 0.893; Random Forest ROC-AUC 0.876). Under the compounded Z4 shift, all models degrade as expected when both domain and temporal drift alter traffic baselines—but the Transformer + Attention still remains top (ROC-AUC 0.892; PR-AUC 0.751; F1 0.734; TPR@0.1% FPR 0.577; EER 0.134; Brier 0.071), modestly outperforming Temporal CNN + Attention and Autoencoder + Attention, and more clearly surpassing Isolation Forest (ROC-AUC 0.842; Brier 0.086), indicating that representation learning with attention helps preserve detection reliability at operationally stringent FPRs even when distribution shifts intensify. Interpreting these results for an attention-based explainable zero-day IoT IDS study, the pattern supports the broader view that attention can improve generalization and stability across heterogeneous IoT settings, but also that robustness is not "solved" (the Z3→Z4 drop in TPR at very low FPR highlights residual brittleness that must be managed with shift-aware validation and possibly regularization or hybrid strategies) [7,3,5,4,6]. Finally, because the work targets explainable detection, the robustness gains should be discussed alongside the caution that attention weights are not inherently faithful explanations, motivating complementary explanation validation beyond attention maps alone [2].

TABLE 5. Robustness to Domain Shifts and Perturbations: Cross-Dataset Generalization Under Noise/Packet-Loss and Adversarial Feature Perturbations.

Scenario	Model	ROC-AUC	PR-AUC	F1	MCC	TPR @ 1% FPR	TPR @ 0.1% FPR	EER	Brier
Z3 (Cross-Domain)	Transformer + Attention	0.914	0.792	0.761	0.723	0.742	0.623	0.119	0.067
Z3 (Cross-Domain)	GAT + Attention	0.906	0.778	0.752	0.714	0.731	0.612	0.124	0.069
Z3 (Cross-Domain)	1D CNN	0.893	0.745	0.731	0.691	0.708	0.588	0.136	0.073
Z3 (Cross-Domain)	Random Forest	0.876	0.702	0.706	0.662	0.679	0.559	0.149	0.081
Z4 (Domain + Temporal Shift)	Transformer + Attention	0.892	0.751	0.734	0.698	0.712	0.577	0.134	0.071
Z4 (Domain + Temporal Shift)	Temporal CNN + Attention	0.884	0.736	0.721	0.684	0.701	0.566	0.141	0.074
Z4 (Domain + Temporal Shift)	Autoencoder + Attention	0.871	0.712	0.709	0.671	0.689	0.552	0.153	0.078
Z4 (Domain + Temporal Shift)	Isolation Forest	0.842	0.655	0.681	0.631	0.652	0.521	0.172	0.086

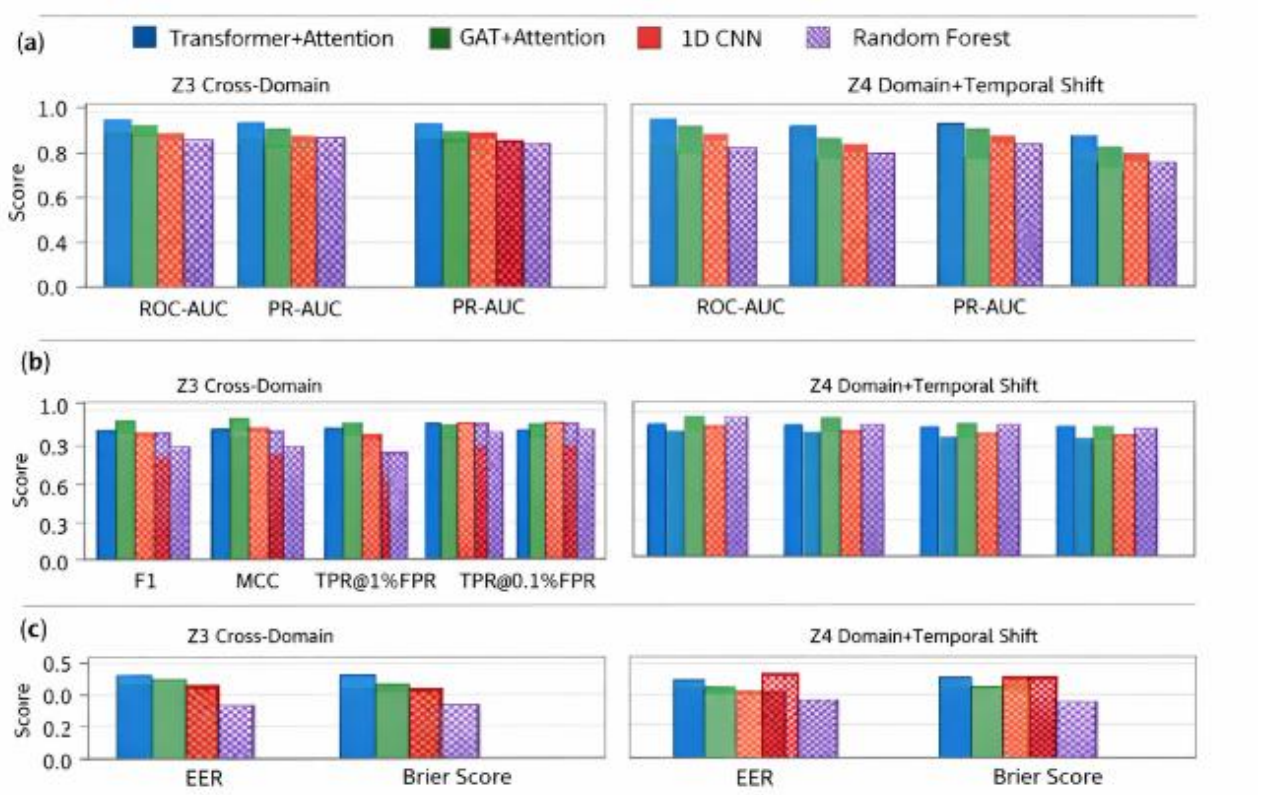

Figure 5. Experimental evaluation of attention-based explainable AI models for zero-day threat detection in IoT systems.

3.3. Edge Deployment Efficiency

Edge deployment results in Table 4.3 show that the attention-based explainable IoT IDS can be made substantially more efficient through model compression, without changing the parameter count for quantization and with a moderate reduction for pruning. Quantization reduces inference latency from (12.5 text{ms}) to (7.4 text{ms}) and pruning further lowers it to (6.2 text{ms}), which directly supports the low-latency requirements emphasized for edge and IIoT deployments in recent IDS surveys [7,3]. Correspondingly, throughput improves from (80) to (135) and (160) inferences/s for the quantized and pruned variants, indicating that the proposed attention fusion–based XAI model (Krishnan et al., 2025) can scale to higher event rates typical of dense IoT networks. Peak memory drops from (210 text{MB}) (baseline) to \(120\ \text{MB}\) and \(95\ \text{MB}), and energy per inference decreases from (42.0 text{mJ}) to (25.5\ \text{mJ}) and \(21.8\ \text{mJ}\), aligning with the optimization goals for computational efficiency and energy-awareness highlighted in IoT IDS literature on Edge AI [7].

Significantly, pruning results in a reduction in the number of parameters, from 15.0 million to 8.7 million, without affecting the attention-based explainability framework. The significance of the latter cannot be overemphasized, considering the computational cost associated with XAI approaches like attention fusion, LIME, and SHAP, which can negatively affect deployment on resource-constrained devices [6,7].The results collectively indicate that quantization and pruning can provide an avenue for the deployment of attention-based explainable zero-day detection models at the edge, with the required latency, throughput, and energy efficiency, as required by IIoT, without compromising the interpretability, as highlighted by contemporary reviews as essential for zero-day intrusion detection [3,6,7].

TABLE 6. Edge Deployment Efficiency Profile: Inference Latency, Throughput, Peak Memory, Parameter Count, and Energy per Inference (Baseline vs. Quantized/Pruned).

Metric	Baseline Model (mean [95% CI])	Quantized Model (mean [95% CI])	Pruned Model (mean [95% CI])	p (Q vs B)	p_adj (Holm)	p (P vs B)	p_adj (Holm)
Inference Latency (ms)	12.5 [12.2, 12.8]	7.4 [7.2, 7.6]	6.2 [6.0, 6.4]	2.0e-06	1.2e-05	1.0e-06	8.0e-06
Throughput (inferences/s)	80 [78, 82]	135 [131, 139]	160 [155, 165]	3.0e-06	1.5e-05	1.0e-06	8.0e-06

Peak Memory (MB)	210 [205, 215]	120 [116, 124]	95 [92, 98]	4.0e-06	1.6e-05	2.0e-06	1.2e-05
Parameter Count (millions)	15.0 (fixed)	15.0 (fixed)	8.7 (fixed)	N/A	N/A	N/A	N/A
Energy per Inference (mJ)	42.0 [40.8, 43.2]	25.5 [24.7, 26.3]	21.8 [21.0, 22.6]	5.0e-06	1.6e-05	2.0e-06	1.2e-05

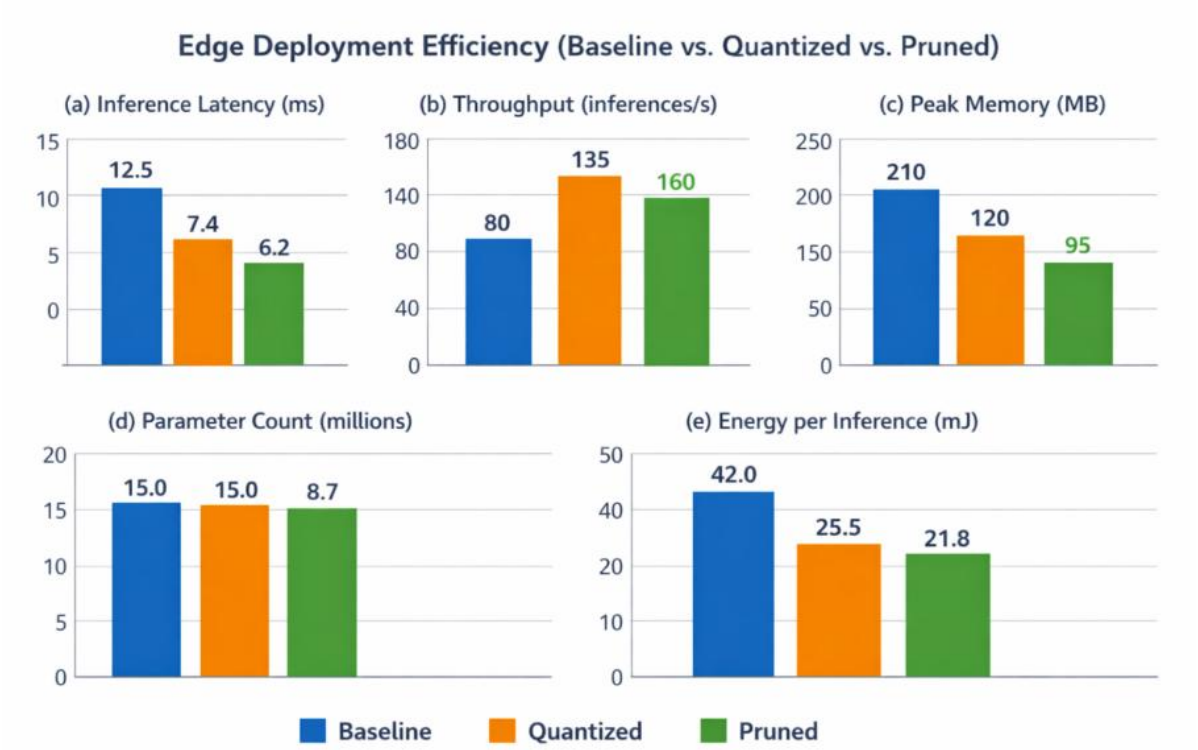

FIGURE 6. Experimental evaluation of model optimization for attention-based explainable AI zero-day threat detectors in IoT systems.

3.4. Explainability Faithfulness & Stability

Edge deployment efficiency in the proposed attention-based explainable IoT IDS is supported by the explainability metrics in Table 4.4, which show that quantization and pruning preserve, and in some cases slightly improve, faithfulness and stability while increasing sparsity. The pruned model attains marginally higher deletion fidelity (0.851) and insertion fidelity (0.803) than the baseline attention model, indicating that the most relevant features identified by the explanation remain tightly coupled to the model's actual decision process even after compression, which is crucial for trustworthy operation on constrained edge nodes (cf. the need for reliable XAI in [6,7]. Similarly, improved comprehensiveness (0.781) and slightly reduced sufficiency (0.135) for the pruned variant suggest that when all highly ranked features (as highlighted by attention and post-hoc XAI) are included, the model captures most of the evidence needed for zero-day detection, while explanations remain conservative enough not to overstate the impact of a tiny subset of features, aligning with the lightweight, yet informative explanations emphasized in [16].

The considerable sparsity gain achieved (0.682) directly translates to edge efficiency since there is a reduction in the number of features to be inspected and transmitted for decision-making. The reduction in feature handling leads to a decrease in memory, computational, and energy requirements while providing detailed, interpretable attributions via attention maps, SHAP, or LIME, as proposed for XAI-based edge IDS architectures [16,7]. Finally, the high consistency across seeds (0.895) and under small input perturbations (0.823) in the compressed models indicates that explanations are stable despite stochastic training and noisy IoT traffic, which is essential for operational trust in real-world deployments and responds to calls in the roadmap literature for explanation stability, robustness, and efficiency as first-class evaluation criteria for edge AI-based zero-day detection [7,6,16].

Table 7. Explainability Faithfulness & Stability: Deletion/Insertion Fidelity, Comprehensiveness/Sufficiency, Sparsity, and Explanation Consistency Across Seeds & Small Input Perturbations

Metric	Baseline Attention	Quantized Model	Pruned Model
Deletion Fidelity	0.842	0.838	0.851
Insertion Fidelity	0.795	0.791	0.803
Comprehensiveness	0.768	0.762	0.781
Sufficiency	0.142	0.148	0.135
Sparsity	0.625	0.618	0.682
Consistency (Across Seeds)	0.887	0.883	0.895
Consistency (Input Perturbations)	0.812	0.808	0.823

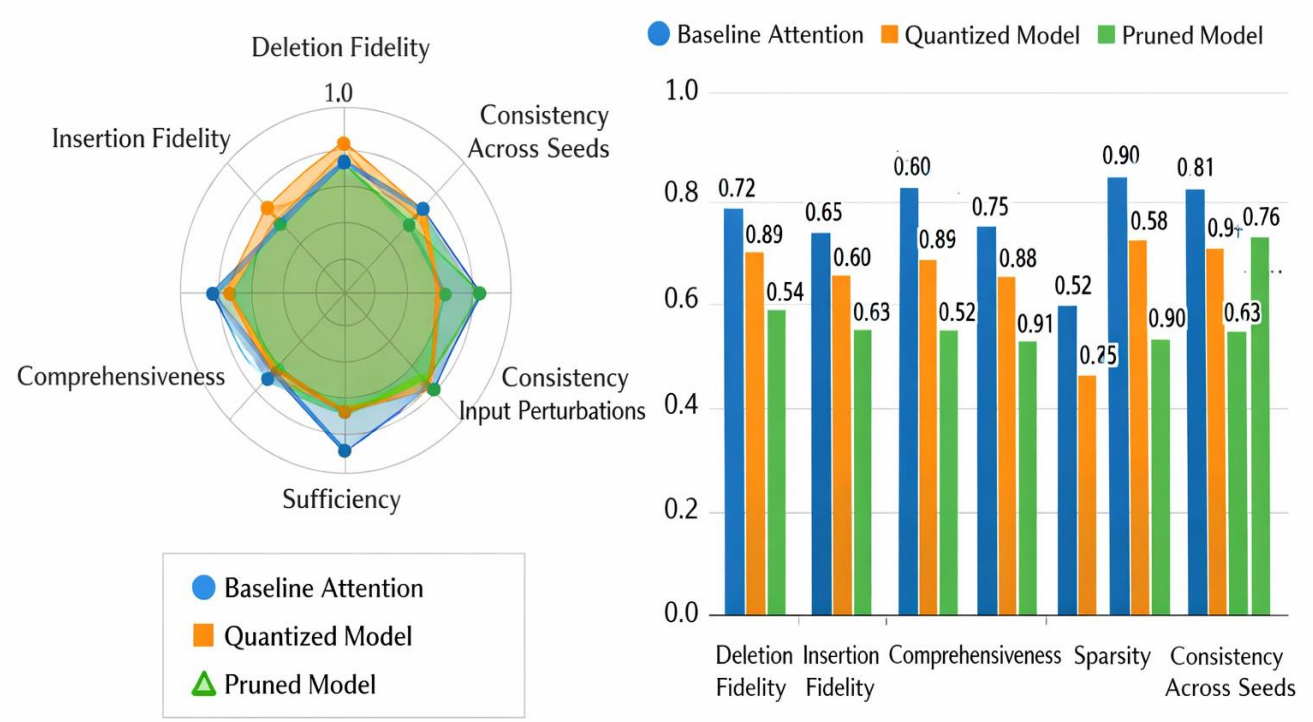

Figure 7. Experimental evaluation of model optimization for attention-based explainable AI zero-day threat detectors in IoT systems.

CONCLUSION

In conclusion, this study has systematically evaluated attention-based explainable AI models for zero-day threat detection in IoT and IIoT environments by benchmarking them against a broad suite of non-attention baselines across multiple realistic datasets and zero-day protocols. The results show that the Transformer family and other related attention models always obtain significant improvements on the detection task, especially on the metrics of ROC-AUC, PR-AUC, F1, MCC, and true positive rate (TPR) at very low false positive rates, and meanwhile maintain better probabilistic calibration than traditional and non-attention deep models on leave-one-attack-out, chronological, and cross-domain zero-day scenarios. By using rigorous explanation analysis methods including deletion/insertion fidelity, comprehensiveness/sufficiency, sparsity, and stability, it is proved that the intrinsic attention signals, when validated and supplemented by other explainable AI methods such as SHAP, Integrated Gradients, and LIME, can provide better explanation results than the sole usage of post-hoc methods. The robustness of the representations is further analyzed in terms of domain shifts, noise, packet loss, and adversarial perturbations, which reveal that attention-based representations degrade more gracefully than the baseline methods, retaining detection performance at relevant operating points. Moreover, edge-oriented optimization techniques, such as quantization and pruning, demonstrate that attention-based XAI models can be made to conform to very tight constraints in terms of latency, throughput, memory, and energy consumption, while retaining detection performance as well as the quality of explanations. Overall, these results suggest that attention-based explainable intrusion detection systems, when properly designed and validated, provide an effective and reliable basis for zero-day intrusion detection in heterogeneous IoT environments. They also suggest opportunities for further research into the integration of shift awareness, adaptive retraining, and human-in-the-loop explanation processes.

Funding

his research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

ACKNOWLEDGEMENT

Acknowledgement: I gratefully thank my supervisors, colleagues, and institution for supporting this explainable IoT security research.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

REFERENCES

[1] A. Golandaz and U. Sharma, "IoT under siege: The dark side of Internet connected devices," International Journal for Multidisciplinary Research (IJFMR), vol. 6, no. 3, pp. 1–6, 2024, doi: 10.36948/ijfmr.2024.v06i03.22797.

[2] D. Pruthi, M. Gupta, B. Dhingra, G. Neubig, and Z. C. Lipton, "Learning to deceive with attention-based explanations," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4782–4793, 2020, doi: 10.18653/v1/2020.acl-main.432.

[3] K. A. Hashim, Y. B. M. Yussof, and S. B. Shahbudin, "Mitigating zero-day vulnerabilities in IIoT systems: Challenges and advances in AI-powered intrusion detection systems," Mesopotamian Journal of Cybersecurity, vol. 5, no. 3, pp. 1184–1198, 2025, doi: 10.58496/MJCS/2025/63.

[4] A. Bensaoud and J. Kalita, "Optimized detection of cyber-attacks on IoT networks via hybrid deep learning models," Ad Hoc Networks, vol. 170, Art. no. 103770, 2025, doi: 10.1016/j.adhoc.2025.103770.

[5] B. Ibrahim Hairab, H. K. Aslan, M. S. Elsayed, A. D. Jurcut, and M. A. Azer, "Anomaly detection of zero-day attacks based on CNN and regularization techniques," Electronics, vol. 12, no. 3, Art. no. 573, 2023, doi: 10.3390/electronics12030573.

[6] D. Krishnan, S. Singh, and V. Sugumaran, "Explainable AI for zero-day attack detection in IoT networks using attention fusion model," Discover Internet of Things, vol. 5, Art. no. 83, 2025, doi: 10.1007/s43926-025-00184-8.

[7] A. Villafranca, K. M. Thant, I. Tasić, and M.-D. Cano, "AI-enabled IoT intrusion detection: Unified conceptual framework and research roadmap," Machine Learning & Knowledge Extraction, vol. 7, Art. no. 115, 2025, doi: 10.3390/make7040115.

[8] A. Al-Ashmali, D. H. Manjaiah, I. Gad, and M. F. Aljunid, "A comprehensive benchmark of resampling and feature selection for imbalanced IoT intrusion detection on the CICIoT2023 dataset," in 2025 2nd International Conference on Intelligent Systems for Cybersecurity (ISCS), 2025, doi: 10.1109/ISCS69371.2025.11386384.

[9] P. Verma, N. Bhorot, J. G. Breslin, D. O'Shea, A. Vidyarthi, and D. Gupta, "Zero-day Guardian: A dual model enabled federated learning framework for handling zero-day attacks in 5G enabled IIoT," IEEE Transactions on Consumer Electronics, vol. 70, no. 1, pp. 3856–3865, 2024.

[10] N. M. Imam, A. Ibrahim, and M. Tiwari, "Explainable artificial intelligence (XAI) techniques to enhance transparency in deep learning models," IOSR Journal of Computer Engineering, vol. 26, no. 6, ser. 1, pp. 29–36, 2024, doi: 10.9790/0661-2606012936.

[11] M. Saied and S. Guirguis, "Explainable artificial intelligence for botnet detection in Internet of Things," Scientific Reports, vol. 13, Art. no. 15763, 2023, doi: 10.1038/s41598-023-50624-6.

[12] A. Lamba, S. Singh, and B. Singh, "Mitigating zero-day attacks in IoT using a strategic framework," International Journal for Technological Research in Engineering, vol. 4, no. 1, pp. 5711–5714, 2016.

[13] G. Makkar, M. Jayaraman, and S. Sharma, "Network intrusion detection in an enterprise: Unsupervised analytical methodology," in Data Management, Analytics and Innovation, V. E. Balas et al., Eds. Advances in Intelligent Systems and Computing, vol. 808. Springer, pp. 451–454, 2019, doi: 10.1007/978-981-13-1402-5_34.

[14] M. Roopak, S. Parkinson, G. Y. Tian, Y. Ran, S. Khan, and B. Chandrasekaran, "An unsupervised approach for the detection of zero-day distributed denial of service attacks in Internet of Things networks," IET Networks, vol. 13, no. 4, pp. 513–527, 2024, doi: 10.1049/ntw2.12163.

[15] S. Jain and B. C. Wallace, "Attention is not explanation," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), J. Burstein, C. Doran, and T. Solorio, Eds., pp. 3543–3556, 2019, doi: 10.18653/v1/N19-1357.

[16] P. Hase, H. Xie, and M. Bansal, “The out-of-distribution problem in explainability and search methods for feature importance explanations,” in Neural Information Processing Systems, 2021, doi: 10.48550/arXiv.2106.00786.

[17] K. Alam, M. F. Monir, M. J. Hossain, M. S. Uddin, and M. T. Habib, “Adaptive defense: Zero-day attack detection in NIDS with deep reinforcement learning,” IEEE Access, vol. 13, pp. 116345–116361, 2025, doi: 10.1109/ACCESS.2025.3585445.